A Need for Data-centric Semantics-based Infrastructure for e-Science

Marian Bubak, Tomasz Gubała, Maciej Malawski,
Piotr Nowakowski, and Tomasz Szepieniec

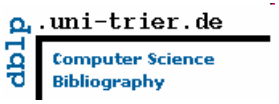
Institute of Computer Science, AGH
Academic Computer Centre – CYFRONET

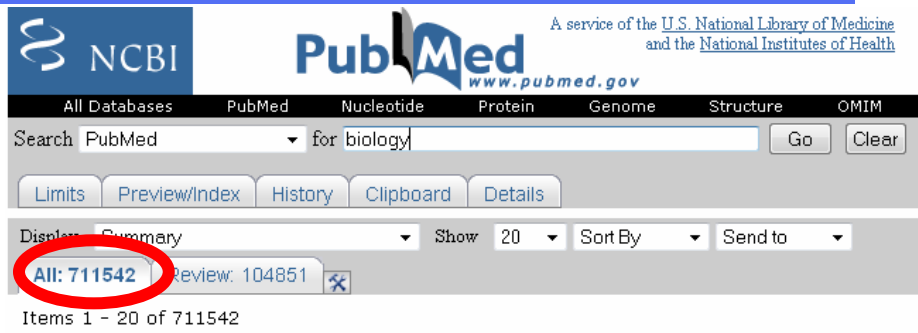
Outline

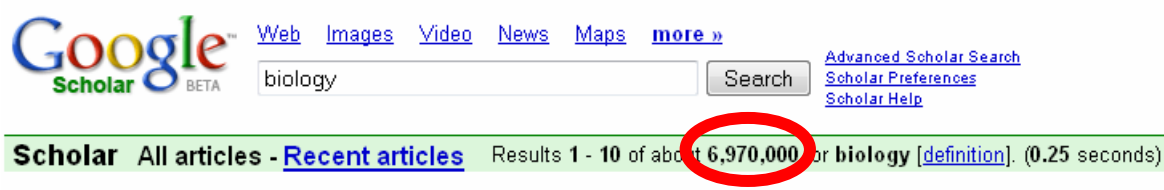
- Open Science and Science 2.0
- Scientific Infrastructures
- Virtual laboratories and access to computing Grid resources
- Concept of Dataneum infrastructure
- Ontology-based annotation framework
- Data Web platform and authorization model

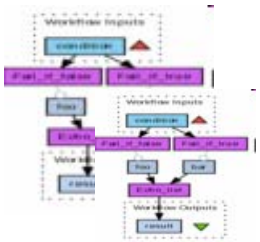
Scientific Publications, Data, Experiments

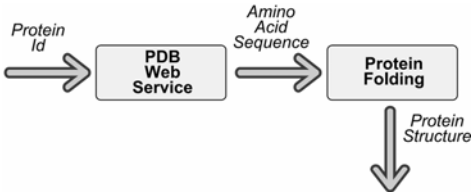
- Large number of publications makes research difficult
  - Computer Science: DBLP contains more than 2^{20} = 1,048,576 publications
- Huge amount of scientific data consumed and produced by e-Science
  - HEP
  - Bioinformatics
- Plentitude of scientific software: jobs, workflows, services, components, scripts, experiment plans...
- Need to link publications with primary data (experimental data, algorithms, software, workflows)
- Reproducible experiments, provenance in e-Science

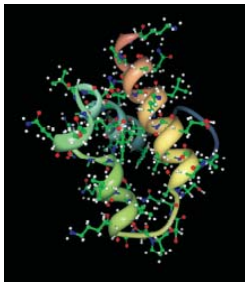








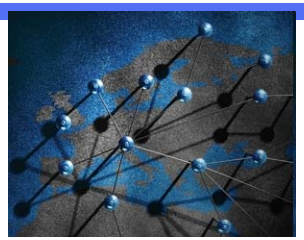




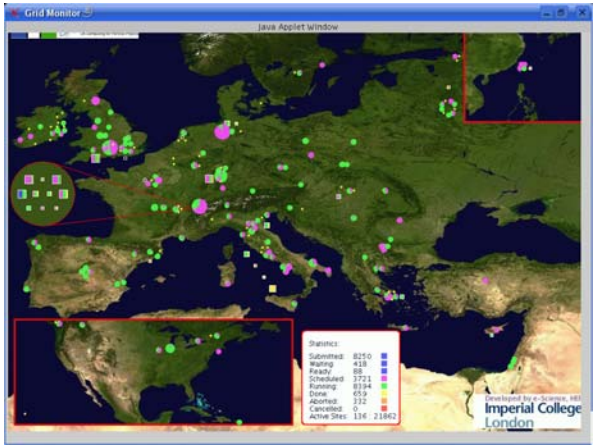
Open Science & Science 2.0

- New means of scientific communication:
  - Wikis, blogs,
  - collaborative web 2.0 technologies
- New way of performing science: e-science, in-silico experiments, exploratory applications
- Democratization of science
- Increasing role of openness

European Scientific Infrastructures

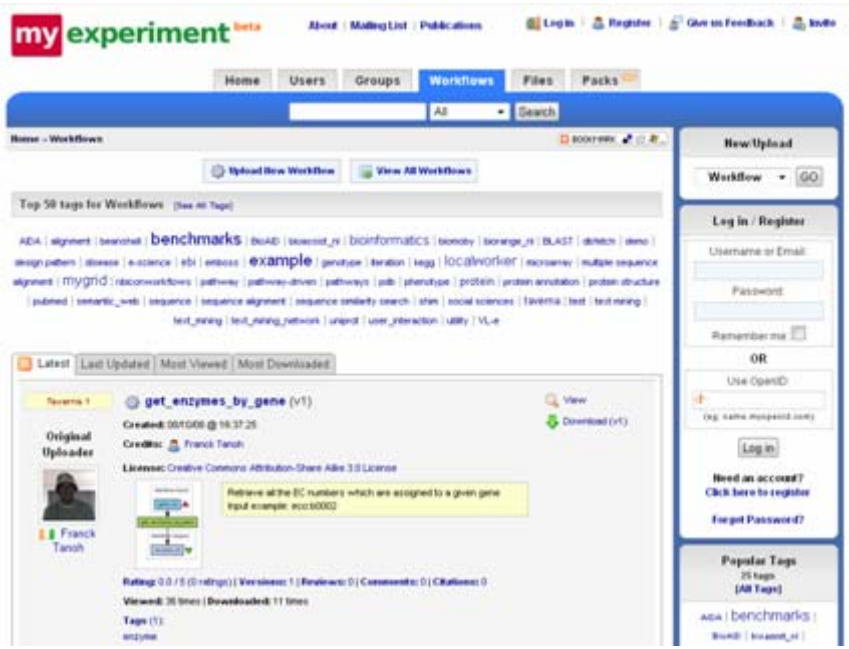


- Grid Infrastructures
  - EGI (EGEE, NGIs) – see CGW'08 Monday programme ☺
  - Reaching 100000 CPU cores
- Supercomputing
  - DEISA



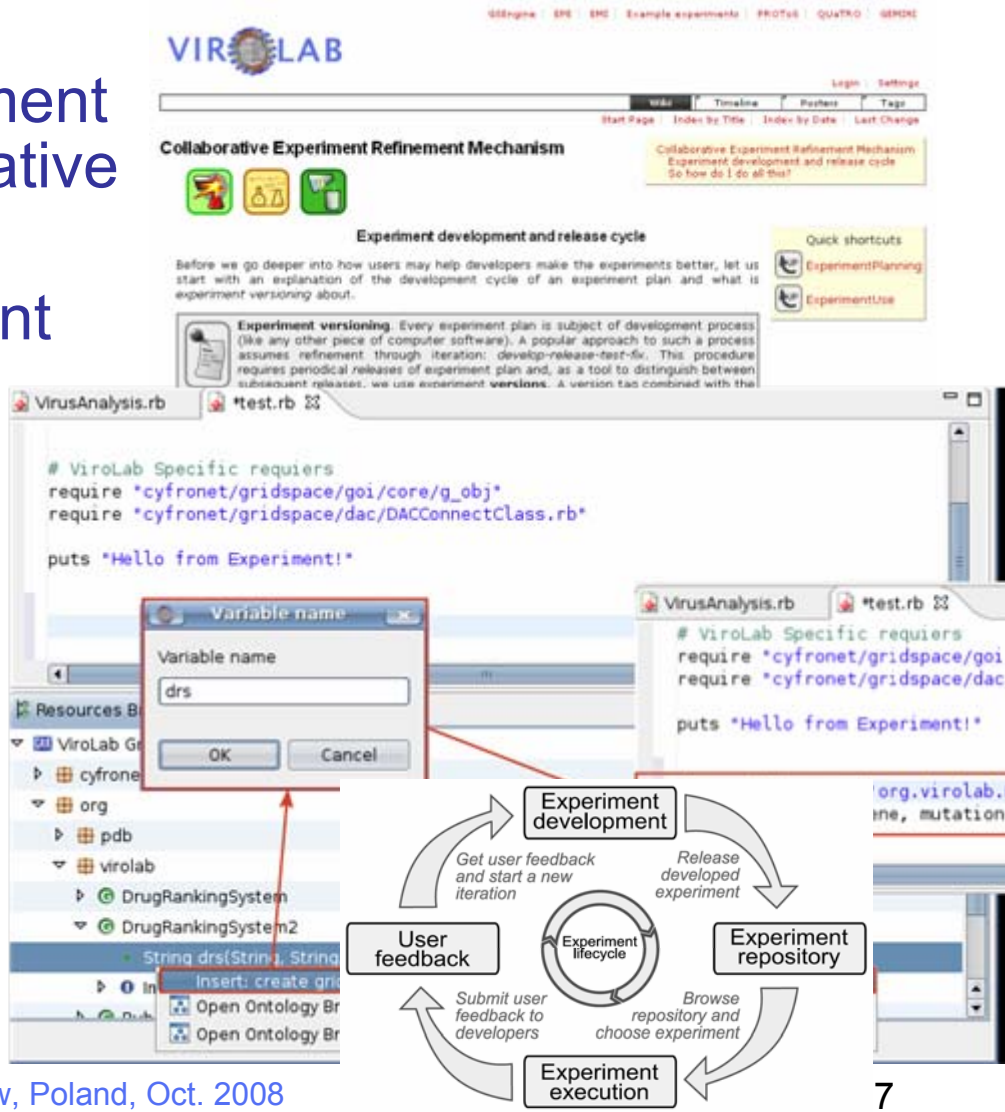
myExperiment

- Social networking site for scientists
  - Publish, share and reuse
  - Workflows, digital objects, collections (packs)
  - Credits, attributions, licensing
- Community
  - Over 1204 users, 99 groups, 459 workflows, 130 files and 36 packs



ViroLab Virtual Laboratory

- Environment for development and execution of collaborative applications
- Scripting-based experiment plans (Ruby)
- Experiment Planning Environment
- Experiment Management Interface
- Experiment Repository
- Result Management
- Access to wide range of middleware (Grid, Web)



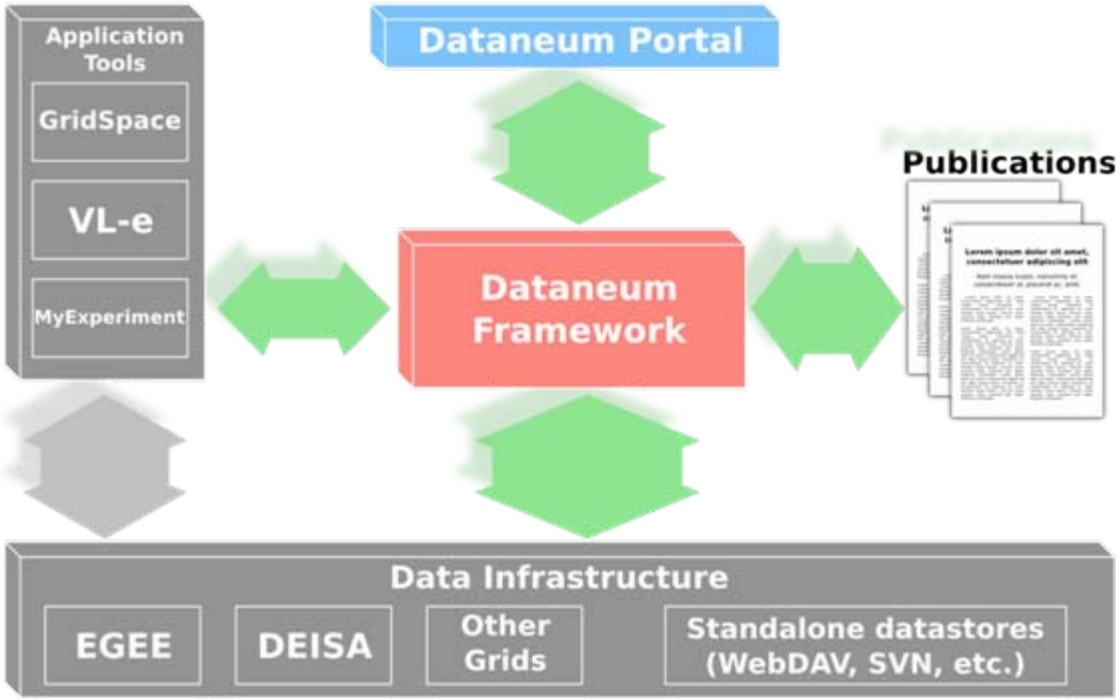
Objectives of our Research

- A uniform and generalized methodology
  - to describe and reference scientific data (including algorithms),
  - enriching scientific publications with a data context,
  - including support for: sharing, reuse, validation, linking with other types of data.
- A web platform for researchers to publish, annotate and access data, according to the defined methodology
- To deploy services for proper access to the data, including grid software repositories
- An authorization model
  - dynamic grouping of researchers
  - customizing privileges freely,
  - adjustable to the scope and state of research

A Concept of the new Infrastructure

- Integrated infrastructure for:
  - Authoring
  - Publishing
  - Managing
  - Sharing
  - Referencing
  - Accessing
  - Reusing
  - Annotating

  scientific data,



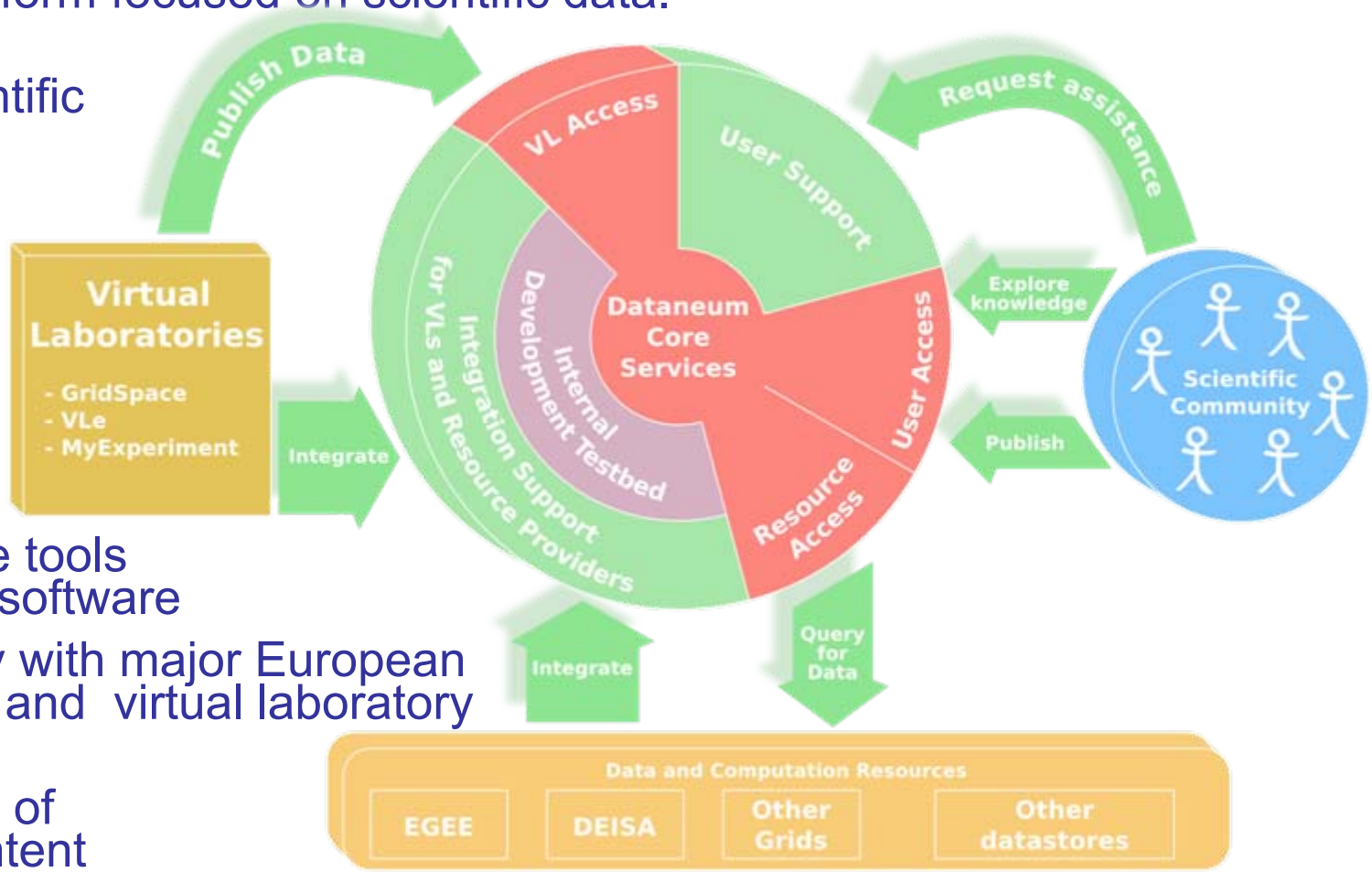
- Framework could reference data stored on storages belonging to EGEE, DEISA, NGS, AlmereGrid, etc.
- Extensions for major European Virtual Laboratory frameworks: Taverna, myExperiment, VL-e, ViroLab GridSpace,

Ontology-based Annotation Framework

- Objectives
  - To define an annotation framework supporting typed annotation of arbitrary objects including experiments, results, users, annotations etc
  - Tools and services supporting the creation, storage, update and publication of annotations
  - Compatibility of the developed framework with existing and emerging mechanisms and frameworks for metadata
- Foundation
  - S-OGSA, Research Objects
  - Multiple types: free text, ontologies, folksonomy tags
  - principles of Linked Data (http://linkeddata.org/)

Data Web Platform and Authorization Model

- Web 2.0 platform focused on scientific data.
- Environment helping scientific community building



- Collaborative tools for scientific software
- Compatibility with major European grid projects and virtual laboratory frameworks
- Visualization of scientific content

Conclusions

- Web 2.0 changes the way scientists work
- Infrastructures are there, but they need to be more user friendly
- Virtual laboratories help users run their experiments
- There is a need to combine these efforts to create an integrated infrastructure
- Building on experience of myGrid, ViroLab, Taverna will bring us closer to the solution
- Partners:
  - Universiteit van Amsterdam
  - HLRS Stuttgart
  - ACC CYFRONET-AGH Kraków
  - University of Manchester
  - Genias-Benelux
  - University Magna Graecia of Catanzaro